

Spécification du moteur de vérification

Moteur de vérification d'Hallucide (src/hallucide/). Document établi depuis le code source. Chaque section correspond à un module réel.

Principe

Le moteur est un pipeline déterministe : le modèle de langage décompose la question et met en forme la réponse, mais le verdict sur chaque affirmation vient toujours de la source officielle. Aucun modèle ne juge sa propre fidélité.

Les sept étapes du pipeline (une par module)

Étape	Module	Rôle
1. Décomposition	<code>_1_decomposition/</code>	La question est découpée en intentions atomiques par le LLM (Claude par défaut, Mistral et Gemini interchangeables). Les routes structurées (article de code, commissions d'un député) court-circuitent le LLM : chemin 100 % déterministe.
2. Couverture	<code>_2_coverage/</code>	Contrôle que chaque intention de la question est bien traitée. Une couverture insuffisante fait monter le risque.
3. Récupération	<code>_3_retrieval/</code>	L'orchestrateur appelle lui-même les sources : serveur MCP tricoteuses (codes consolidés, questions parlementaires, mandats et commissions), data.gouv.fr (données tabulaires), fichiers CSV/ZIP. Le LLM ne choisit jamais la source.
4. Vérification	<code>_4_verification/</code>	Chaque affirmation est comparée mot pour mot au passage officiel récupéré, après normalisation (casse, ponctuation de bord, virgule décimale). Le cycle de vie du texte est re-contrôlé : un texte abrogé n'est jamais authentifié. Une couche lexicale déterministe (sans embeddings, sans modèle, sans réseau) signale les reformulations douteuses ; elle peut seulement augmenter le risque, jamais authentifier.

Étape	Module	Rôle
5. Triage	<code>_5_triage/</code>	Plancher de risque incontournable : référence inférée, troncature, pertinence non garantie, couverture insuffisante, sélection ambiguë ou requête partagée entre plusieurs intentions forcent le risque à « élevé ». L'appelant ne peut pas le contourner.
6. Validation	<code>_6_validation/</code>	Tout résultat à risque élevé est marqué « non publiable en l'état » et exige une décision humaine explicite, capturée dans un registre. Le mode document garantit qu'une synthèse ne peut pas omettre silencieusement une partie du texte.
7. Audit	<code>_7_audit/</code>	Chaque vérification produit un journal de conformité rejouable. Il ne contient jamais la question posée ni l'identité de l'utilisateur, garde-fou vérifié par assertion.

Les statuts d'une affirmation

Statut	Sens	Condition établie par le code
AUTHENTIFIÉ	Citation sûre et invocable	Présente mot pour mot dans une source opposable en vigueur
DONNÉE_TRACÉE	Donnée chiffrée sûre	Retrouvée à la cellule exacte d'une donnée publique officielle
CITÉ_NON_OPPOSABLE	Vrai mais non invocable	Verbatim exact, mais source non opposable (débat, exposé des motifs) ou texte abrogé
INTERPRÉTATION	Reformulation à valider	Recouvrement lexical suffisant et toutes les négations et valeurs chiffrées ancrées dans la source
NON_AUTHENTIFIÉ	Potentielle hallucination	Non retrouvée dans la source ; bloquée

Cas particulier : quand aucune affirmation vérifiable ne peut être produite, le moteur répond `NO_ANSWER`. Il se tait plutôt que d'inventer.

Garanties structurelles

- L'opposabilité dérive du type de document et de son cycle de vie, jamais d'un paramètre de l'appelant ni du modèle.
- Le risque ne descend jamais sous « élevé » une fois qu'une condition déterministe l'a déclenché.

- Un document (note, synthèse) n'a jamais de statut agrégé : chaque affirmation garde le sien.
- Une intention à risque élevé n'est jamais publiée sans décision humaine journalisée.
- Rien n'est simulé : sans clé API ou sans source joignable, le système le dit explicitement.

Interfaces

- `Hallucide.ask(question)` : point d'entrée unique du moteur.
- `POST /resolve` : détection de route déterministe (regex), renvoie la route et les champs préremplis.
- `POST /ask` : pipeline complet, renvoie les intentions, leurs affirmations, statuts, sources et journaux ; la couche de présentation y ajoute un score 0-100 et une bande de couleur, traduction fixe du statut moteur.

Tests

220 tests automatisés couvrent le pipeline, dont les invariants ci-dessus, le mode document et les routes structurées. Deux scénarios ont été joués en direct contre les sources réelles : une question à prémisse fausse n'a produit aucune invention (NO_ANSWER), et un passage authentique mais hors sujet a été bloqué par le plancher de risque.